DOCUMENT RESUME

ED 333 038 TM 016 561

AUTHOR Bradley, Raymond T.; And Others
TITLE The Effects of Number of Score Groups and Score Group Size on the Mantel-Haenszel Alpha.
PUB DATE Apr 91
NOTE 54p.; A previous version of this paper was presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 4-6, 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Comparative Analysis; Ethnicity; *Groups; *Licensing Examinations (Professions); Minority Groups; Multiple Choice Tests; Sample Size; Scores; *Scoring; *Test Interpretation; Test Results
IDENTIFIERS *Alpha Coefficient; *Mantel Haenszel Procedure; Post Hoc Methods

ABSTRACT

Test data from a large licensure examination were used to investigate the impact of the number of score groups and score group size on the Mantel-Haenszel (MH) alpha statistic. Two kinds of quasi-experimental studies were conducted: (1) baseline MH analyses in which the number of score groups and the ethnicity of each randomly equivalent reference-focal group pair were held constant while score group size was systematically reduced from 50 cases to 5 cases per score group; and (2) majority group versus minority group MH analyses in which both the number of score groups and ethnicity varied as the minimum score group size was reduced from 50 cases to 10 cases per score group. The responses of 68,458 candidates to 299 multiple-choice items in a licensure examination administered in 1988 were analyzed in both studies. The results from both kinds of studies show that alpha behavior remained remarkably stable as score group size was reduced. A post hoc MH analysis extending these conditions to their logical extreme--a minimum score group size of one case and a maximum score interval of a single raw score point--also show that the alpha statistic remained stable. Fifteen tables and two graphs illustrate the analyses; and an appendix of two tables summarizes the post hoc analyses.
(Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

RAYMOND T. BRADLEY

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED333038

The Effects of Number of Score Groups and Score Group Size on the Mantel-Haenszel Alpha[1]

Raymond T. Bradley

Anne R. Fitzpatrick

Robert C. Sykes

CTB Macmillan/McGraw-Hill

April, 1991

A previous version of this paper was presented at the annual meeting of the National Council on Measurement in Education, Chicago, April, 1991.

[1] We would like to express our gratitude to Tom Gambill who wrote the computer program for the Mantel-Haenszel procedure, ran (with Mary Ferrier's assistance) the analyses reported in this paper, and contributed thoughtful suggestions to the research reported here.

TM016561

BEST COPY AVAILABLE

The Effects of Number of Score Groups and Score Group Size on the Mantel-Haenszel Alpha

ABSTRACT

Test data from a large licensure examination were used to investigate the impact of the number of score groups and score group size on the Mantel-Haenszel alpha statistic. Two kinds of quasi-experimental studies were conducted: baseline MH analyses in which the number of score groups and the ethnicity of each randomly equivalent reference-focal group pair was held constant while score group size was systematically reduced from 50 cases to 5 cases per score group; and majority group versus minority group MH analyses in which both the number of score groups and ethnicity varied as the minimum score group size was reduced from 50 cases to 10 cases per score group. The results from both kinds of studies showed that alpha behavior remained remarkably stable as score group size was reduced. A post hoc MH analysis extending these conditions to their logical extreme -- a minimum score group size of one case and a maximum score interval of a single raw score point -- also found that alpha remained stable.

The Effects of Number of Score Groups and Score Group Size on the Mantel-Haenszel Alpha

Originally developed for the analysis of matched groups comparisons in bio-medical research, the Mantel-Haenszel (MH) procedure (Mantel & Haenszel, 1959) has been shown to be a powerful and effective method of detecting potential item bias that is uniform (Hambleton & Rogers, 1989; Hills, 1989; Holland & Thayer, 1986). As described by Holland and Thayer (1986), the procedure compares the item performance of a reference (majority) group and a focal (minority) group that have been matched on ability. For this comparison, examinees are assigned to score groups usually on the basis of their total test score, and the examinees in each score group are classified in terms of whether or not they answered the item correctly. A contingency table is then constructed for each reference/focal group pair that is matched on ability. An odds ratio (called alpha) of the likelihood of success on the item is computed to compare the performance of the reference and focal group and summed across all of the matched reference/focal group pairs. Alpha (α) can vary from zero to infinity. When $\alpha = 1.00$ the odds of success on the item are equal for the reference and focal groups, and it is inferred that the item has no potential item bias. To the degree that the odds of success on the item are greater for the reference group than for the focal group, alpha will be greater than 1.00.

Current literature on the use of the MH procedure has not provided definitive answers to two questions relevant both to a more complete theoretical understanding of alpha behavior and also of import for the practical application of the procedure. The first question concerns the number of score groups that should be used in matching the abilities of examinees. Raju, Bode, and Larsen (1989) recommended that at least four score groups be used, whereas Wright (1986) indicated that 61 score groups were more desirable than six. Hambleton and Rogers (1989) used each raw score in a 75-item test to define a score group.

The second question concerns the number of examinees required in each score group (cf. Baghi & Ferrara, 1989; Hambleton & Rogers, 1989; Linacre, 1988; Scheuneman & Bleistein, 1989). This is a particularly important matter, since the number of examinees in a minority group can often be quite small. Scheuneman and Bleistein (1989), Hambleton and Rogers (1989), and Baghi and Ferrara (1989) recommended that "large" samples be used, but they did not specifically say how many examinees are needed. Linacre (1988) suggested combining scores groups that contain few examinees but did not specify a minimum number of examinees that should comprise a score group.

The larger concern underlying both of these issues is the question of the conditions (and operational criteria) for adequate measurement of the matching variable. This is important because the MH procedure requires that differences in ability between the reference and focal groups are controlled within the

test population. Alpha is then estimated on the assumption that ability is held constant. It is not clear, though, in the literature, the size and number of score groups required to achieve adequate measurement of the control variable. Consequently it is not clear how much "error" is introduced into the alpha estimate when non-optimal measurement of the controlling variable occurs. Yet, under normal testing conditions, the number of score groups and the number of cases in a score group are operationally related: the minimum number of cases set for score group construction will determine the number of score groups that can be developed for a given analysis. The setting for both of these parameters will determine, in turn, the minimum population size required for inclusion of a given focal group. Thus, the relationship between these two factors, in a given testing situation, will influence not only the quality of measurement achieved on the controlling variable, but also the number of focal groups that can be included for analysis.

With the two objectives of contributing to an enhanced understanding of alpha behavior and to develop guidelines for practitioners, the present research investigated the impact of the number of score groups and score group size on the behavior of the MH alpha using test data from a large licensure examination. Two studies were conducted using quasi-experimental research designs. First, baseline MH analyses were conducted in which the ethnicity of each reference-focal group pair was the same, that is, held constant. For each of the three ethnic

populations investigated in the baseline studies, candidate identity, the number of score intervals, the interval size, and the region of the score scale included in the analyses all were held constant while the number of cases per score group was systematically varied. The second study was designed to assess the generalizability of the baseline results. In this study, the same test data were used and the ethnicity of the reference and focal groups differed. MH analyses were conducted using various score group sizes and various numbers of score intervals.

Baseline Studies

Description of Sample and Test Data

The data analyzed consisted of 68,458 candidate responses to 299 multiple-choice items in a licensure examination administered in 1988. A total of 47,573 of the candidates classified themselves in Ethnic Group 1, 6,486 classified themselves in Ethnic Group 2, 5,466 classified themselves in Ethnic Group 3, and 2,004 classified themselves in Ethnic Group 4. In addition, 1,014 candidates classified themselves in Ethnic Group 5, 486 classified themselves in Ethnic Group 6, and 307 classified themselves in Ethnic Group 7. Finally, 726 candidates classified themselves as "Other," and the remaining 4,396 candidates did not assign themselves to an ethnic category. Figure 1 shows plots of the raw score distributions for the Ethnic Groups 1 through 7, and Table 1 shows the raw score means and standard deviations for these seven ethnic groups. Since Ethnic Groups 1, 2, and 3 were the largest, they were arbitrarily selected for the

baseline analyses. Candidates in the remaining four ethnic groups, candidates who classified themselves as "Other," and those who did not assign themselves to an ethnic category were not included in the baseline analyses.

Method

Variables. The term Score Group Size is used to refer to the number of cases in a score group. Interval Size is used to refer to the number of raw score points in the score range covered by a score group.

Procedure. Two baseline studies were conducted to determine the effect of Score Group Size on the MH alpha when ethnicity and the number of score groups were held constant. To conduct these studies members of the same ethnic group were randomly assigned to either the reference group or the focal group. For example, the members of Ethnic Group 1 were randomly assigned either to the reference group, designated Ethnic Group 1a, or to the focal group, designated Ethnic Group 1b. Similarly, the members of Ethnic Group 2 and Ethnic Group 3 were randomly assigned to Ethnic Groups 2a and 2b and Ethnic Groups 3a and 3b, respectively. By making the ethnicity of the reference and focal groups in a pair the same, a baseline showing the impact of changes in Score Group Size on the MH alpha could be established with the confounding effects of ethnicity removed.

In both baseline studies, Score Group Size for both the reference and focal groups initially were set equal to 50 by randomly selecting 50 reference group members and 50 focal group

members for each score group. In Baseline Study I Score Group Size for the focal group then was systematically decreased from 50 to 25, 20, 15, 10 and 5 randomly selected cases. In Baseline Study II Score Group Size for both the reference and focal groups was decreased by these amounts. In both studies, candidate identity was fixed so that the groups of 50 candidates initially selected for each score group were used as the source for the randomly selected candidates used for the smaller score groups of 25, 20, 15, 10 and 5 candidates. By using this procedure to select the smaller score groups that had less than 50 cases, the effects of sampling error were minimized in the analyses involving these groups.

In both baseline studies the number of score groups that were used was held constant for the analyses conducted for each reference-focal group pair, but the number of score groups varied over the reference-focal group pairs. Development of the score groups was done separately for each reference-focal group pair. For a pair, this development began at the median raw score of the focal group using an automated procedure. This median score was treated as the first possible score group with an interval size equal to one raw score point. The focal group and the reference group were checked to establish whether or not they had at least 50 cases in this first score group. If either group failed to meet this criterion value, the raw score range covered by the first score group was expanded until there were at least 50 cases in both the reference and focal groups, with the limit that the

raw score range for the score group could not exceed 10 raw score points. Subsequent score groups were constructed on either side of the first score group using the same procedure.

As a result of imposing these constraints, it was found that Ethnic Groups 1a and 1b could be divided into 100 score groups, Ethnic Groups 2a and 2b into 45 score groups, and Ethnic Groups 3a and 3b into 37 score groups. Also, for the comparisons conducted for each reference-focal group pair some cases at the tails of the score distributions were excluded from the analyses because fewer than 50 cases within a 10-point raw score interval were available for the reference and focal groups. This was done in order to minimize any confounding effects on the MH alpha that result from poorly matched cases in the tails.

Table 2 shows the analyzed cases and the cases excluded from the analyses, as well as the raw score ranges of the analyzed and excluded cases. Table 3 shows the Interval Sizes of the score groups created for each reference/focal group pair. The data in Table 3 demonstrate that although the maximum Interval Size was 10 raw score points, in all three types of reference/focal group comparisons, the vast majority of Score Groups spanned only one or two raw score points.

Results

Baseline Study I: Reducing the Focal Group Score Size. Tables 4 and 5 show the results of MH analyses performed on the three pairs of reference and focal groups as Score Group Size for the focal group was reduced systematically from 50 to 5 randomly

selected cases. When ethnicity is held constant, no DIF should be detected when the MH procedure is applied. Consistent with this expectation, the results of Baseline Study I shown in Table 4 indicate that the mean alpha for each of the three types of reference/focal group comparisons involving 50 cases per score group for both the reference and focal groups was very close to or equal to 1.00.[2] The standard deviations of the alphas estimated in these comparisons also were near 0.00. When Score Group Size for the focal group was reduced from 50 cases to 5 cases, the mean alpha increased by no more than 0.009, and the standard deviations of the alphas increased by no more than .099. For all three sets of comparisons, the greatest change in the alphas occurred when the Score Group Size for the focal groups was decreased from 10 to 5 cases. Table 5 shows that the MH analyses conducted in all three sets of comparisons produced only one item with an alpha value greater than 1.50 as Score Group Size for each of the focal groups was decreased from 50 cases to 10 cases. The one item with an alpha value greater than 1.50 was detected in the comparison of Ethnic Groups 3a vs. 3b, when the Score Group Size for the focal group was 10 cases. This item had an alpha of 1.527. Indeed, even when Score Group Size for the focal group dropped to 5 cases for the comparison involving

---

[2] Due to the ordinal nature of the alpha (odds ratio) scale, the mean alphas can not be added or subtracted to evaluate the impact of changes in Score Group Size. However, the means are reported and discussed in this research because they can be of assistance in detecting an effect of Score Group Size.

Ethnic Groups 1a vs. 1b, the maximum alpha value obtained was 1.337. However, when Score Group Size was reduced to 5 cases for the focal groups in Ethnic Groups 2b and 3b, four items with alpha values between 1.40 and 1.59 were found for each of these two groups.

Baseline Study II: Reducing Reference and Focal Score Group Size. Tables 6 and 7 show the results of MH analyses conducted for the three types of reference/focal group pairs when Score Group Size for both the reference and focal groups was systematically decreased from 50 to 5 randomly selected cases. The results in Table 6 indicate that as Score Group Size was reduced from 50 to 5 cases for both the reference and focal groups, the mean alpha increased by no more than 0.020 and the standard deviation of the alphas increased by no more than 0.162. In all three sets of comparisons, the means and standard deviations of the alphas increased the most when Score Group Size for the reference and focal groups was reduced from 10 to 5 cases.

As Table 7 shows, the increases in the means and standard deviations were due to the slight tendency for the alpha values to increase as Score Group Size decreased. For example, although there were no items with alphas $\geq$1.50 when 50 cases per score group were analyzed, when these cases were decreased to 10 cases, one item with an alpha $\geq$1.50 was found for the EG 1a vs. 1b comparison, four items for the EG 2a vs. 2b comparison, and five items with alphas $\geq$1.50 were found for the EG 3a vs. 3b comparison. When Score Group Size was dropped to 5 cases, again

one item with an alpha $\geq$1.50 was found for the EG 1a vs. 1b comparison, 11 items with alphas $\geq$1.50 were found for the EG 2a vs. 2b comparison, and 12 items with alphas $\geq$1.50 were found for the EG 3a vs. 3b comparison.

Discussion

The findings from the baseline studies suggest that, within the conditions examined, the MH alpha appears relatively robust with respect to decreases in Score Group Size from 50 to 10 cases. The first baseline study demonstrates that applying the MH procedure in a reference-focal group comparison with 10 cases per score group in lieu of 50 cases per interval for a focal group effected only very slight changes in the MH alpha results that were obtained. The findings from this study also suggest that this conclusion holds regardless of whether one is using 100 score intervals or 37 Score intervals.

The findings from the second baseline study indicate that applying the MH procedure in a reference-focal group comparison with 10 cases per score interval in lieu of 50 cases per interval for both the reference and focal groups also effects very little change in the MH alpha. Again, this conclusion appears to hold regardless of whether one is using 100 score intervals or 37 score intervals.

Majority vs. Minority Group Comparisons

Given the promising results of the baseline studies, an additional study was undertaken to extend the type of research conducted in the baseline studies to the circumstance in which

the reference and focal groups differed in ethnicity. In this study, sets of MH comparisons between reference (majority) and focal (minority) groups were conducted as Score Group Size was reduced from a minimum of 50 cases to a minimum of ten cases.

The baseline studies had indicated that Score Group Size could be reduced to 10 cases with little change in the MH alpha when the ethnicity of the reference and focal groups was the same. This finding held whether the reference and focal groups were similar or different in size.

It was important to establish whether or not these findings would also be obtained when the reference and focal groups differed in ethnicity, particularly when the focal (minority) groups were quite small in size. Including as many ethnic groups as possible in a bias analysis is usually a priority of a testing program. If the MH alpha is stable only when score group sizes of 50 or more cases are used, then small minority groups would lack enough cases to be analyzed. On the other hand, if it could be established that small score groups of 10 cases produced MH alphas that were as stable as the MH alphas produced when the score groups were large, then small ethnic groups could be included in MH analyses.

Description of Sample and Test Data

The data analyzed were drawn from the same pool of 68,458 candidates used for the Baseline Studies I and II.

Method

The second set of studies involved using the MH procedure to compare the item performance of reference (majority) and focal (minority) groups that differed in ethnicity as Score Group Size was decreased. Specifically, comparisons were made between the item performance of Ethnic Group 1 (the majority group) and the item performance of each of six minority groups, Ethnic Groups 2 and 3, which were used in Baseline Studies I and II, plus Ethnic Groups 4, 5, 6, and 7, which were not included in the baseline studies. The analyses were extended to Ethnic Groups 4 through 7 in the interest of establishing the impact of decreasing Score Group Size for all possible ethnic groups of interest to the client for whom the licensure examination was developed. Ethnic Groups 4 through 7 were considerably smaller in size than the other ethnic groups. As noted above, including these smaller ethnic groups in the analyses presented an opportunity to examine the extent to which any findings that were obtained could be extended to minority groups that varied considerably in size.

In a departure from the procedure used for the baseline studies, where Score Group Size was completely controlled, only minimum case counts for Score Group Size were established for this second set of studies. This was done to accommodate the client's need to use all available cases in the analyses.

Procedure. A total of five sets of reference/focal group comparisons were conducted, where each set consisted of a number of majority/minority group comparisons and a baseline comparison.

To compose the majority group and the groups for the baseline comparison, members of Ethnic Group 1, which was the largest ethnic group, were randomly assigned to one of two groups of approximately equal size. These groups were called Ethnic Group 1a and Ethnic Group 1b. Ethnic Group 1a was used as the majority group in the majority/minority group comparisons. To provide baseline results, the item performance of Ethnic Group 1a also was compared to that of Ethnic Group 1b.

For each set of comparisons, a minimum case count for Score Group Size was established. For the first set, this Minimum Score Group Size was 50 cases for each score group, for the second set this minimum was 25 cases for each score group, and for the third, fourth, and fifth sets, this minimum was 20 cases, 15 cases, and 10 cases, respectively.

To permit comparisons of results across the comparisons within a set, the score groups used in these comparisons were fixed, so that all comparisons within the set were conducted using the same score groups. In general, the number of score groups established for a set was determined by considering the score distribution of the smallest ethnic group, considering the score distributions of the other ethnic groups, and determining how many score groups could be composed that met both the case count requirement for Minimum Score Group Size and the additional requirement that the Interval Size for a score group should not exceed 10 raw score points. If the smallest ethnic group could not be divided into at least nine score groups, it was excluded

from the analysis and the score distribution of the next smallest ethnic group was used to determine the number of score groups.

A procedure like that used to compose the score groups for the baseline analyses was used for the present study. More specifically, development of the score groups began at the median raw score of the smallest ethnic group using an automated procedure. This median score was treated as the first possible score group with an interval size equal to one raw score point. The smallest ethnic group as well as the others were checked to establish whether or not they had sufficient cases to meet the criterion value for Minimum Score Group Size for the particular comparison set. If any group failed to meet this criterion value, the raw score range covered by the first score group was expanded until there were sufficient cases in all ethnic groups, with the limit that the raw score range for the score group could not exceed 10 raw score points. Subsequent score groups were constructed on either side of the first score group using the same procedure. As noted previously, after all score groups were constructed, they were counted. If the smallest ethnic group did not contain nine score groups, as noted previously, this ethnic group was dropped from the analyses and the construction of score groups was begun again using the second smallest ethnic group.

As was done in the baseline studies, cases in the tails of the score distributions for the ethnic groups were excluded from the analyses when the requisite number of cases could not be collected within an interval of 10 raw score points. This was

done in order to minimize any confounding effects on the MH alpha due to poorly matched cases in these tails.

Tables 8, 9, and 10 show the results of applying these procedures. Specifically, in Table 8 are shown the criteria used to compose each set of comparisons and the impact of imposing these criteria on the number of score groups and the range of raw scores available for analyses in each set. In Table 9 are shown for each comparison the mean Score Group Size and the number of cases that either were analyzed or excluded from the analyses. Finally, in Table 10 are shown the Interval Sizes of the score groups that were composed for each set of analyses.

Table 8 shows that when the Minimum Score Group Size was $\geq$50, 13 score groups having raw scores that ranged from 141 to 201 could be composed for the first set of analyses. Because nine score groups that met the requirements for Minimum Score Group Size and Interval Size could not be constructed for Ethnic Groups 6 and 7, these two ethnic groups were excluded from the first set of comparisons. When the Minimum Score Group Size was dropped to $\geq$25 for the second set of analyses, Table 8 shows that 14 score groups having raw scores that ranged from 134 to 218 could be composed for the second set of analyses. Ethnic Group 7 again was excluded from this set of analyses because it did not have sufficient cases to meet the requirements. When the Minimum Score Group Size was dropped further to $\geq$20 cases for the third set of analyses, only nine Score Groups with a raw score range of 160 to 213 could be composed for the analyses. The number of

score groups and raw score range declined for this set of analyses because Ethnic Group 7, the smallest ethnic group, was used as the basis for developing the score groups. In this case, the priority of evaluating as many ethnic groups as possible was purchased at the price of reducing the number of score groups and the raw score range that could be analyzed. As the Minimum Score Group Size was subsequently decreased to ≥15 cases and ≥10 cases for the fourth and fifth sets of analyses, the number of score groups and the raw score ranges included in the analyses again increased, because the Minimum Score Group Size was now small enough to permit a greater number of score groups and an expanded range of raw scores to be included in the analyses.

Table 9 shows that the actual score group size for all but the smallest ethnic groups included in the analyses substantially exceeded the Minimum Score Group Size established for each set. The smallest ethnic groups that were included in each set of analyses had mean score group sizes that were very close to the Minimum Score Group Sizes that were established for each set.

With respect to the number of cases that were excluded from the analyses, Table 9 shows that for all ethnic groups analyzed in Sets 1 and 2 this number decreased as the Minimum Score Group Size declined from ≥50 cases to ≥25 cases. When the Minimum Score Group Size declined to ≥20 cases in Set 3, Ethnic Group 7 was included in the analysis for the first time. In this set the number of excluded cases increased because the size of Ethnic Group 7 restricted the range of raw scores and the number of

score groups that could be included in the analyses. As the Minimum Score Group Size further declined to ≥15 cases and to ≥10 cases, the number of excluded cases again decreased.

Table 10 shows that of the score groups that were analyzed, at least 50% of the score groups in each set spanned a raw score interval of five points or less. Only in Set 2 was the maximum Interval Size of 10 raw score points used for any score group included in the analyses.

Results

Table 11 presents the results of the MH analyses conducted for each of the five sets of analyses. In the interest of clarity, these results are organized by type of reference/focal group comparison.

Consideration of the baseline results given in Table 11 suggests that the magnitude of potential item bias varied considerably over the six majority/minority comparisons that were conducted. For the baseline comparison involving Ethnic Group 1a vs. 1b, the mean alpha was very close to 1.00, the standard deviation was close to 0.00, and the maximum alpha obtained was 1.14. These values provided the expected evidence that potential bias was not evident in the item performance in the baseline groups. Slightly higher means, standard deviations, and maximum alpha values were observed for the comparisons of Ethnic Group 1a vs. 3, Ethnic Group 1a vs. 4, and Ethnic Group 1a vs. 7. These results suggest that potential bias did not affect much of the item performance of these three minority groups. In contrast,

the comparisons of Ethnic Group 1a vs. 6 and, in particular, the comparisons of Ethnic Group 1a vs. 2 and Ethnic Group 1a vs. 5, produced alphas with values considerably higher than the baseline results. These higher values suggest that potential bias affected to a greater extent the item performance of these three minority groups.

Despite the presence of potential item bias in some of the comparisons, the results in Table 11 suggest that there was a general tendency for the mean alpha and the standard deviation of the alphas to decline as the Minimum Score Group Size decreases, provided that this decrease in Minimum Score Group Size was accompanied by an increase in the number of score groups. This can be seen by examining first the results for Sets 1 and 2 and then the results for Sets 3, 4, and 5.

For Sets 1 and 2, as the Minimum Score Group Size is reduced from $\geq$50 cases to $\geq$25 cases, the number of score groups increased from 13 to 14. The results for Sets 1 and 2 show that the mean alpha and the standard deviation of the alphas remained stable or declined for all comparisons included in these sets.

When compared with the results for Sets 1 and 2, the results for Set 3 showed slight increases in alpha means and standard deviations. Recall that for this set, the Minimum Score Group Size was $\geq$20, but the number of score groups dropped to nine because the smallest ethnic group, Ethnic Group 7, was included in the analyses for the first time. Thus, although the Minimum Score Group Size was decreased from Set 2 to Set 3, the pattern

of declining means and standard deviations did not hold in the presence of the fewer score groups used in Set 3.

When the results for Set 3 are compared to the results for sets 4 and 5, the general pattern of decreasing alpha means and standard deviations once again occurred. Across these sets, while Minimum Score Group Size decreased from ≥20 cases, to ≥15 cases, and then to ≥10 cases, the number of score groups increased from nine groups, to 12 groups, to 19 groups. Hence, the reduction in Minimum Score Group Size was accompanied by increases in the numbers of score groups. The few exceptions to this rule occurred for the baseline comparison and the comparisons of Ethnic Group 1a vs. 4 and Ethnic Group 1a vs. 6 when the Minimum Score Group Size decreased from ≥15 cases to ≥10 cases. In these instances, there were extremely slight increases in the means and standard deviations of the alpha values.

Table 12 shows the distribution of alpha values for each of six types of reference/focal group comparisons under the five Minimum Score Group Size conditions. Consistent with the findings shown in Table 11, the data in Table 12 show that the baseline comparison and the comparisons involving Ethnic Group 1a vs. 3, Ethnic Group 1a vs. 4, and Ethnic Group 1a vs. 7 produced the most items with low alpha values. Considerably more items with higher alpha values were identified from the comparison of Ethnic Group 1a vs. 6 and particularly from the comparisons of Ethnic Group 1a vs. 2 and Ethnic Group 1a vs. 5.

When the results in Table 12 are examined, it appears that as the Minimum Score Group Size was reduced from ≥50 cases to ≥10 cases the item alpha values generally were stable, particularly at the high end of the alpha scale. This finding is most clearly seen in Table 13, where the alpha values given in Table 12 are grouped into larger categories. In Table 13, it is useful to consider first the comparisons that produced considerable numbers of items with alphas ≥3.000. These were the comparisons of Ethnic Group 1a vs. 2 and Ethnic Group 1a vs. 5. When the Minimum Score Group Size was ≥50 cases, 11 items and 17 items were detected for these two comparisons, respectively. As the Minimum Score Group Size was decreased from ≥50 cases to ≥10 cases, the number of items with alphas ≥3.000 decreased slightly in these two comparisons. When the alpha range of 2.000-2.999 is considered, very slight increases and decreases were observed. As the Minimum Score Group Size decreased, more notable shifts in the distributions of item alphas were found in the 1.500-1.999 range and particularly the 1.100-1.499 range.

Although correlations between the 299 alpha values obtained at the five levels of Minimum Score Group Size were not computed, spot checks of these values for a sample set of items indicated that these values changed very little as Minimum Score Group Size decreased. Thus the shifts in the distributions that were observed appear to have resulted simply from small changes in the alpha values computed for the items.

Discussion

The purpose of the second study was to investigate the extent to which the findings from the first study, involving only baseline comparisons, could be extended to the circumstance in which the ethnicity of the reference and focal groups differed in ethnicity. The number of score groups used in the second study was considerably smaller than the number used in the baseline studies because the second study included some minority groups that were quite small in size; their size restricted the number of score groups available that met the requirements imposed in the study.

The findings from the second study appeared to indicate that the MH alphas remained fairly stable, and even declined in some instances, as the minimum number of cases per score group decreased from $\geq$50 cases to $\geq$10 cases. This finding of stable alphas is consistent with the results of the baseline studies conducted in the first investigation, which indicated that systematic decreases in score group size from 50 to 10 cases produced no marked changes in the MH alphas.

In the second study, decreases in the case count requirements per score group were generally accompanied by increases in the number of score groups analyzed. The one exception to this rule occurred when the minimum case counts were dropped from $\geq$25 cases to $\geq$20 cases and the smallest ethnic group was incorporated in the analyses. In this instance, the decline in cases per score group was accompanied by a decrease in the

number of score groups analyzed. This concomitant decrease of both the minimum score group size and the number of score groups appeared to produce a slight increase in the item alphas. This finding suggests that using fewer score groups can inflate the alpha results for items.

Given the evidence of relatively stable results obtained from the majority/minority and baseline comparisons, post hoc analyses were conducted that extended the research to its logical conclusion by investigating the behavior of the MH alpha under the conditions of a Minimum Score Group Size of one case and a Maximum Interval Size equal to one raw score point. Using the same test data and the same majority and minority groups examined in the second study, this sixth set of analyses was conducted to explore the consequences of using the smallest possible interval size, which has been recommended by Hambleton and Rogers (1989), and the smallest possible score group size. Of interest was the question of whether these extreme conditions would reveal a point at which the alphas no longer remained relatively stable.

This sixth set of analyses involved the same procedures for constructing score groups as were used in the previous analyses. Using these procedures, 71 score groups covering the raw score range of 152 to 223 were constructed. Each score group spanned one raw score point. Table 14 shows the mean score group size and the number of analyzed and excluded cases obtained for each comparison.

Table 15 shows the results of the sixth set of analyses, summarizing the distribution of item alpha values across broad ranges on the alpha scale. (A more detailed display of these results is given in Appendix A.) When the results of Table 15 are examined, it appears that these conditions produced only very slight changes in the behavior of the MH alpha when the distribution of alpha values is compared with the results given in Table 13 for the Set 5 analyses described in the second study. In other words, the results of comparisons made with a Minimum Score Group Size of $\geq$10 cases and a maximum Interval Size of 10 raw score points were nearly identical to the results of comparisons made with a Minimum Score Group Size of $\geq$1 case and a maximum Interval Size of one raw score point.

Conclusion

The research reported in this paper was conducted with two objectives in mind. The first was to contribute to an understanding of the influence of score group size and the number of score groups on the behavior of the Mantel-Haenszel alpha. The second was to develop operational guidelines for these two factors for the practitioner who wishes to use the Mantel-Haenszel.

Two studies (a baseline study in which ethnicity was held constant and a study of ethnically diverse majority/minority group pairs), were conducted to investigate the influence of these factors on alpha behavior. With everything held constant

as score group size was systematically varied, the results of the baseline study showed that alpha behavior remained relatively stable as the score group size was reduced from 50 cases down to 10 cases per score group, and that this result held both when only the focal score group size was reduced and also when both the reference and focal score groups' size were reduced concurrently. At a score group size of five cases, though, there was evidence of increased alpha instability.

The second study was conducted to determine whether these results would be obtained in majority/minority group comparisons involving different ethnic groups. Here, though, due to the operational constraints of the testing program, the tight experimental controls established for the baseline study had to be relaxed. Instead of an exact score group size, as was employed in the baseline study, a minimum score group size was used. The results from the second study were comparable to the baseline study in that alpha behavior was found to be remarkably stable as the minimum score group size was reduced from 50 cases per score group to a minimum of 10 cases. There was virtually no change in alpha as the minimum score group size was reduced.

This result prompted an additional post hoc analysis extending the study to its logical conclusion to investigate alpha behavior under the most extreme conditions of a minimum score group size of one case and a maximum score interval of a single raw score point. These results, too, revealed that alpha

behavior remained stable when the alpha values were compared to those obtained at higher minimum score group sizes.

By way of closing, there are two lines of implications that flow from this research. The first concerns some basic theoretical questions which should be addressed in future research regarding alpha behavior. How might we account for such stable alpha behavior as score group size is reduced from 50 cases, through ten cases, down to a single case? One potentially plausible explanation is that reducing score group size permits improved measurement of the matching variable, subsequently allowing both a more precise matching and matching across a larger region of the raw score scale. However, the slight change in alpha that was evident in the results of this research was a systematic tendency for a reduction of alphas with higher values.

A second plausible explanation may lie with the weighting of the alpha statistic. The notion here is that the weighting feature of the MH alpha is able to sufficiently compensate for the sparsity of scores in certain regions of the scale score and thus may enable valid DIF estimates without introducing grouping effects.

A second line of implications are more practical in nature. One important implication, insofar as these results are found to obtain more generally, is that bias information could be obtained on a broad range of ethnic groups. This is a direct consequence of the reduced minimum score group size that would enable the procedure to be extended to smaller focal groups for which,

previously, bias information was not obtainable. Taken in conjunction with the results of the post hoc analysis, it would appear that a minimum score group size of ≥ 1 case and a maximum interval size of one raw score point would provide an optimal operational circumstance for the practitioner who desires bias information from as many focal groups as possible. Under these conditions, the MH alpha would, ipso facto, be independent of any particular grouping scheme and consequently would more likely be generalizable across, or independent of, the distributions of ability. Further study of the behavior of the MH alpha under these conditions when the majority and minority distributions do and do not differ is merited.

A final question, to be investigated by future research, concerns the extent to which the results hold for examinations of a much shorter length. This is an important issue since the stability of alpha behavior observed in this research may be attributable in part to the robustness with which the underlying ability construct is measured. In the present research ability was measured across 299 items. It will be essential to see how alpha behaves in tests of a much smaller length (say 40-50 items) as the factors we have investigated here are manipulated.

References

Baghi, H., & Ferrara, S. (1989). A comparison of IRT, Delta plot, and Mantel-Haenszel techniques for detecting differential item functioning across subpopulations in the Maryland Test of Citizenship Skills. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased item items: Comparison of IRT area and Mantel-Haenszel methods. Applied Measurement in Education, 2, 313-324.

Hills, J. (1989). Screening for potentially biased items in testing programs. Educational Measurement: Issues and Practice, 8, 5-11.

Holland, P. W., & Thayer, D. T. (1986). Differential item performance and the Mantel-Haenszel procedure. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Linacre, J. M. (1988). The practical realization of the standard error of the Mantel-Haenszel statistic. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Raju, N. S., Bode, R. K., & Larsen, V. S. (1989). An empirical assessment of the Mantel-Haenszel statistic for studying differential item performance. Applied Measurement in Education, 2, 1-13.

Scheuneman, J. D., & Bleistein, C. A. (1989). A consumer's guide to statistics for identifying differential item functioning. Applied Measurement in Education, 3, 255-275.

Wright, D. H. (1986). An empirical comparison of the Mantel-Haenszel and standardization methods of detecting differential item performance. (Statistical Report No. SR-86-99). Princeton, NJ: Educational Testing Service.

Table 1

Raw Score Means and Standard Deviations for
Ethnic Groups 1, 2, 3, 4, 5, 6, and 7

Ethnic Group	N	Mean	Standard Deviation
1	47,573	207.37	21.93
2	6,486	161.94	29.71
3	5,466	183.05	23.24
4	2,004	183.36	29.70
5	1,014	162.12	27.42
6	486	173.25	31.00
7	307	195.42	25.48

Table 2

Baseline Studies I and II

Characteristics of the Analyzed and Excluded Cases by Ethnic Group

Characteristic	Ethnic Group 1	Ethnic Group 2	Ethnic Group 3
Analyzed Cases			
Raw Score Range	152-260	89-215	133-232
N of Cases	47,573	6,486	5,466
Number of Score Groups	100	45	37
Excluded Cases			
Raw Score Range			
Lower tail	0-151	0- 88	0-132
Upper tail	261-299	216-299	233-299
N of Cases			
Lower tail	350	50	132
Upper tail	150	162	60

Table 3

Baseline Studies I and II

Number of Score Groups by Raw Score Interval Size and Reference/Focal Group Comparison

Raw Score Interval Size	Reference/Focal Group Comparison		
	Ethnic Groups 1a vs. 1b	Ethnic Groups 2a vs. 2b	Ethnic Groups 3a vs. 3b
10 R.S. Points	-	1	-
9 R.S. Points	-	-	1
8 R.S. Points	-	-	1
7 R.S. Points	-	1	-
6 R.S. Points	-	2	1
5 R.S. Points	-	2	1
4 R.S. Points	-	2	2
3 R.S. Points	2	8	5
2 R.S. Points	5	27	23
1 R.S. Point	93	2	3
Total No. of Score Groups	100	45	37

Table 4

Baseline Study I

Results of MH Analyses by Reference/Focal (R/F) Group Comparison and R/F Score Group Size

R/F Comparison and R/F Score Group Size	α Mean	α S.D.
EG 1a vs. 1b		
50 cases/50 cases	1.000	.055
50 cases/25 cases	1.000	.065
50 cases/20 cases	1.001	.070
50 cases/15 cases	1.002	.078
50 cases/10 cases	1.003	.091
50 cases/ 5 cases	1.003	.118
EG 2a vs. 2b		
50 cases/50 cases	1.003	.070
50 cases/25 cases	1.003	.083
50 cases/20 cases	1.004	.091
50 cases/15 cases	1.006	.096
50 cases/10 cases	1.008	.113
50 cases/ 5 cases	1.012	.151
EG 3a vs. 3b		
50 cases/50 cases	1.003	.076
50 cases/25 cases	1.004	.094
50 cases/20 cases	1.004	.103
50 cases/15 cases	1.004	.113
50 cases/10 cases	1.003	.126
50 cases/ 5 cases	1.007	.175

Table 5

Baseline Study I

Distribution of Item Alpha Values by Reference/Focal (R/F) Group Comparison and R/F Score Group Size

R/F Comparison and R/F Score Group Size	Alpha Values								
	<1.100	1.100-1.199	1.200-1.299	1.300-1.399	1.400-1.499	1.500-1.599	≥1.600	Min.	Max.
EG 1a vs. 1b									
50 cases/50 cases	290	8	1	0	0	0	0	.751	1.214
50 cases/25 cases	279	18	2	0	0	0	0	.736	1.228
50 cases/20 cases	272	25	2	0	0	0	0	.783	1.268
50 cases/15 cases	263	32	4	0	0	0	0	.825	1.279
50 cases/10 cases	254	39	4	2	0	0	0	.737	1.351
50 cases/ 5 cases	232	51	13	3	0	0	0	.595	1.337
EG 2a vs. 2b									
50 cases/50 cases	267	31	1	0	0	0	0	.798	1.224
50 cases/25 cases	253	42	3	1	0	0	0	.790	1.307
50 cases/20 cases	247	45	6	1	0	0	0	.769	1.313
50 cases/15 cases	250	38	11	0	0	0	0	.801	1.293
50 cases/10 cases	234	52	11	2	1	0	0	.749	1.417
50 cases/ 5 cases	225	43	19	8	2	2	0	.670	1.587
EG 3a vs. 3b									
50 cases/50 cases	271	24	3	1	0	0	0	.738	1.335
50 cases/25 cases	253	36	9	1	0	0	0	.627	1.312
50 cases/20 cases	246	42	8	3	0	0	0	.559	1.335
50 cases/15 cases	238	46	11	4	0	0	0	.570	1.372
50 cases/10 cases	234	47	13	5	0	1	0	.588	1.527
50 cases/ 5 cases	218	32	30	15	3	1	0	.339	1.588

Table 6

Baseline Study II

Results of MH Analyses by Reference-Focal (R/F) Group Comparison and R/F Score Group Size

R/F Comparison and R/F Score Group Size	α Mean	α S.D.
EG 1a vs. 1b		
50 cases/50 cases	1.000	.055
25 cases/25 cases	1.002	.075
20 cases/20 cases	1.004	.084
15 cases/15 cases	1.006	.096
10 cases/10 cases	1.008	.119
5 cases/ 5 cases	1.012	.171
EG 2a vs. 2b		
50 cases/50 cases	1.003	.070
25 cases/25 cases	1.004	.093
20 cases/20 cases	1.005	.103
15 cases/15 cases	1.007	.120
10 cases/10 cases	1.012	.151
5 cases/ 5 cases	1.018	.211
EG 3a vs. 3b		
50 cases/50 cases	1.003	.076
25 cases/25 cases	1.004	.110
20 cases/20 cases	1.007	.123
15 cases/15 cases	1.010	.141
10 cases/10 cases	1.011	.170
5 cases/ 5 cases	1.023	.238

Table 7

Baseline Study II

Distribution of Item Alpha Values by Reference/Focal (R/F) Group Comparison and R/F Score Group Sizes

R/F Comparison and R/F Score Group Size	Alpha Values								
	<1.100	1.100-1.199	1.200-1.299	1.300-1.399	1.400-1.499	1.500-1.599	≥1.600	Min.	Max.
EG 1a vs. 1b									
50 cases/50 cases	290	8	1	0	0	0	0	.751	1.214
25 cases/25 cases	269	29	1	0	0	0	0	.722	1.287
20 cases/20 cases	262	30	6	1	0	0	0	.759	1.335
15 cases/15 cases	252	39	6	1	1	0	0	.755	1.411
10 cases/10 cases	233	49	11	5	0	1	0	.654	1.587
5 cases/ 5 cases	209	44	26	14	5	1	0	.655	1.524
EG 2a vs. 2b									
50 cases/50 cases	267	31	1	0	0	0	0	.798	1.224
25 cases/25 cases	249	43	6	1	0	0	0	.780	1.329
20 cases/20 cases	244	43	11	1	0	0	0	.732	1.309
15 cases/15 cases	232	41	21	5	0	0	0	.699	1.370
10 cases/10 cases	225	45	20	3	2	3	1	.675	1.607
5 cases/ 5 cases	213	48	11	9	7	6	5	.559	1.844
EG 3a vs. 3b									
50 cases/50 cases	271	24	3	1	0	0	0	.738	1.335
25 cases/25 cases	237	49	9	4	0	0	0	.538	1.309
20 cases/20 cases	231	48	14	6	0	0	0	.543	1.393
15 cases/15 cases	218	51	21	5	4	0	0	.586	1.454
10 cases/10 cases	218	37	27	11	1	5	0	.608	1.582
5 cases/ 5 cases	193	36	29	16	13	5	7	.292	1.701

Table 8

Characteristics of Comparison Sets

Set No. and R/F Comparisons	Characteristics of Set: Min. Score Group Size	Max. Interval Size	Raw Score Range	No. of Score Groups
Set 1	≥ 50	10	141-201	13
EG 1a vs. 2				
EG 1a vs. 3				
EG 1a vs. 4				
EG 1a vs. 5				
EG 1a vs. 6*				
EG 1a vs. 7*				
EG 1a vs. 1b				
Set 2	≥ 25	10	134-218	14
EG 1a vs. 2				
EG 1a vs. 3				
EG 1a vs. 4				
EG 1a vs. 5				
EG 1a vs. 6				
EG 1a vs. 7*				
EG 1a vs. 1b				
Set 3	≥ 20	10	160-213	9
EG 1a vs. 2				
EG 1a vs. 3				
EG 1a vs. 4				
EG 1a vs. 5				
EG 1a vs. 6				
EG 1a vs. 7				
EG 1a vs. 1b				
Set 4	≥ 15	10	157-220	12
EG 1a vs. 2				
EG 1a vs. 3				
EG 1a vs. 4				
EG 1a vs. 5				
EG 1a vs. 6				
EG 1a vs. 7				
EG 1a vs. 1b				
Set 5	≥ 10	10	150-221	19
EG 1a vs. 2				
EG 1a vs. 3				
EG 1a vs. 4				
EG 1a vs. 5				
EG 1a vs. 6				
EG 1a vs. 7				
EG 1a vs. 1b				

Note: In the first column, EG is used as an abbreviation for Ethnic Group.

* Indicates a comparison not conducted because the score distributions for the minority group did not meet the requirements for Minimum Score Group Size and Interval Size.

Table 9

Characteristics of Comparisons by Set

Set No. and R/F Comparisons	Mean Score Group Size		Cases Analyzed		Cases Excluded	
	Ref. Group	Focal Group	Ref. Group	Focal Group	Ref. Group	Focal Group
Set 1 (≥50 cases)						
EG 1a vs. 2	706.0	345.8	9,178	4,496	14,657	1,990
EG 1a vs. 3	706.0	315.1	9,178	4,096	14,657	1,370
EG 1a vs. 4	706.0	98.5	9,178	1,283	14,657	723
EG 1a vs. 5	706.0	56.0	9,178	728	14,657	286
EG 1a vs. 6	***	***	***	***	***	***
EG 1a vs. 7	***	***	***	***	***	***
EG 1a vs. 1b	706.0	703.5	9,178	9,146	14,657	14,592
Set 2 (≥25 cases)						
EG 1a vs. 2	1,160.8	371.4	16,251	5,199	7,584	1,287
EG 1a vs. 3	1,160.8	358.6	16,251	5,020	7,584	466
EG 1a vs. 4	1,160.8	120.2	16,251	1,683	7,584	321
EG 1a vs. 5	1,160.8	59.8	16,251	837	7,584	177
EG 1a vs. 6	1,160.8	29.1	16,251	408	7,584	78
EG 1a vs. 7	***	***	***	***	***	***
EG 1a vs. 1b	1,160.8	1,157.8	16,251	16,209	7,584	7,529
Set 3 (≥20 cases)						
EG 1a vs. 2	1,544.8	392.0	13,903	3,528	9,932	2,958
EG 1a vs. 3	1,544.8	469.7	13,903	4,227	9,932	1,239
EG 1a vs. 4	1,544.8	152.0	13,903	1,368	9,932	636
EG 1a vs. 5	1,544.8	61.6	13,903	554	9,932	460
EG 1a vs. 6	1,544.8	33.6	13,903	302	9,932	184
EG 1a vs. 7	1,544.8	23.2	13,903	209	9,932	98
EG 1a vs. 1b	1,544.8	1,541.6	13,903	13,874	9,932	9,864
Set 4 (≥15 cases)						
EG 1a vs. 2	1,399.3	321.3	16,792	3,856	7,043	2,630
EG 1a vs. 3	1,399.3	379.3	16,792	4,552	7,043	914
EG 1a vs. 4	1,399.3	126.9	16,792	4,523	7,043	481
EG 1a vs. 5	1,399.3	50.1	16,792	601	7,043	413
EG 1a vs. 6	1,399.2	28.2	16,792	338	7,043	148
EG 1a vs. 7	1,399.2	23.2	16,792	209	7,043	98
EG 1a vs. 1b	1,399.2	1,396.8	16,792	16,761	7,043	6,997
Set 5 (≥10 cases)						
EG 1a vs. 2	910.2	229.3	17,293	4,356	6,542	2,130
EG 1a vs. 3	910.2	253.7	17,293	4,820	6,542	646
EG 1a vs. 4	910.2	85.7	17,293	1,619	6,542	385
EG 1a vs. 5	910.2	36.7	17,293	697	6,542	317
EG 1a vs. 6	910.2	19.3	17,293	367	6,542	119
EG 1a vs. 7	910.2	13.2	17,293	250	6,542	57
EG 1a vs. 1b	910.2	908.6	17,293	17,260	6,542	6,478

Note: In the first column, EG is used as an abbreviation for Ethnic Group.
*** Indicates a comparison not conducted because the score distributions for the minority group did not meet the requirements for Minimum Score Group Size and Interval Size. minimum case counts within the maximum allowable Interval Size

Table 10

Number of Score Groups by Raw Score Interval Size and Set

Raw Score Interval Size	Set 1	Set 2	Set 3	Set 4	Set 5
10 R.S. Points	-	1	-	-	-
9 R.S. Points	-	-	-	-	1
8 R.S. Points	1	1	-	2	-
7 R.S. Points	1	2	2	1	1
6 R.S. Points	1	2	2	2	1
5 R.S. Points	2	4	2	2	1
4 R.S. Points	5	3	3	4	4
3 R.S. Points	3	1	-	1	5
2 R.S. Points	-	-	-	-	6
1 R.S. Point	-	-	-	-	-
Total No. of Score Groups	13	14	9	12	19

Table 11

Results of MH Analyses by Reference/Focal (R/F) Group Comparison and Set

R/F Comparison and Set No.	Min. Score Group Size	α Mean	α S.D.	α Min.	α Max.
EG 1a vs. 2					
Set 1	≥50	1.207	.865	.155	8.090
Set 2	≥25	1.200	.847	.159	8.051
Set 3	≥20	1.209	.890	.149	8.625
Set 4	≥15	1.204	.880	.154	8.648
Set 5	≥10	1.200	.862	.156	8.364
EG 1a vs. 3					
Set 1	≥50	1.056	.329	.400	2.254
Set 2	≥25	1.056	.324	.411	2.263
Set 3	≥20	1.057	.327	.412	2.294
Set 4	≥15	1.055	.326	.415	2.302
Set 5	≥10	1.055	.325	.410	2.306
EG 1a vs. 4					
Set 1	≥50	1.042	.273	.401	2.014
Set 2	≥25	1.041	.259	.423	2.007
Set 3	≥20	1.040	.255	.453	2.034
Set 4	≥15	1.036	.251	.445	1.978
Set 5	≥10	1.038	.252	.436	1.945
EG 1a vs. 5					
Set 1	≥50	1.305	1.110	.088	9.437
Set 2	≥25	1.291	1.074	.095	9.482
Set 3	≥20	1.304	1.127	.099	9.761
Set 4	≥15	1.298	1.124	.100	9.857
Set 5	≥10	1.297	1.112	.098	9.970
EG 1a vs. 6					
Set 1	≥50	***	***	***	***
Set 2	≥25	1.087	.473	.315	4.304
Set 3	≥20	1.081	.459	.344	3.833
Set 4	≥15	1.081	.456	.348	4.023
Set 5	≥10	1.082	.461	.332	4.163
EG 1a vs. 7					
Set 1	≥50	***	***	***	***
Set 2	≥25	***	***	***	***
Set 3	≥20	1.031	.242	.480	2.061
Set 4	≥15	1.032	.228	.502	2.020
Set 5	≥10	1.025	.222	.532	1.961
EG 1a vs. 1b					
Set 1	≥50	1.000	.036	.871	1.137
Set 2	≥25	1.001	.026	.900	1.086
Set 3	≥20	1.000	.029	.904	1.124
Set 4	≥15	1.000	.026	.922	1.076
Set 5	≥10	1.000	.027	.924	1.093

*** Indicates a comparison not conducted because the score distributions for the minority group did not meet the requirements for Minimum Score Group Size and Interval Size.

BEST COPY AVAILABLE

Table 12

Distribution of Item Alpha Values by
Reference/Focal (R/F) Comparison and Set

R/F Comp. and Set	Alpha Values																				
	<1.100	1.100-1.199	1.200-1.299	1.300-1.399	1.400-1.499	1.500-1.599	1.600-1.699	1.700-1.799	1.800-1.899	1.900-1.999	2.000-2.199	2.200-2.399	2.400-2.599	2.600-2.799	2.800-2.999	3.000-3.199	3.200-2.399	3.400-3.599	3.600-3.799	3.800-3.999	≥4.000
EG 1a vs. 2																					
Set 1	170	16	12	15	12	14	5	5	5	3	12	6	5	4	4	1	1	3	0	2	4
Set 2	173	14	13	15	12	15	5	5	4	2	13	5	4	7	2	0	2	2	0	3	3
Set 3	175	13	10	17	15	10	6	6	3	2	13	6	4	5	4	0	1	1	2	1	5
Set 4	171	16	13	17	11	12	10	2	3	2	13	6	5	4	4	1	1	1	2	0	5
Set 5	170	18	10	19	12	14	4	5	4	2	13	5	5	5	3	0	3	0	2	0	5
EG 1a vs. 3																					
Set 1	185	31	21	17	12	16	6	3	2	0	5	1	0	0	0	0	0	0	0	0	0
Set 2	182	34	16	21	14	15	7	4	0	1	3	2	0	0	0	0	0	0	0	0	0
Set 3	183	29	24	17	18	11	8	1	2	1	4	1	0	0	0	0	0	0	0	0	0
Set 4	183	31	18	23	18	10	5	3	2	1	4	1	0	0	0	0	0	0	0	0	0
Set 5	183	30	21	21	17	11	6	3	1	0	5	1	0	0	0	0	0	0	0	0	0
EG 1a vs. 4																					
Set 1	195	29	24	19	17	3	5	3	1	2	1	0	0	0	0	0	0	0	0	0	0
Set 2	194	32	25	20	11	6	6	2	1	2	0	0	0	0	0	0	0	0	0	0	0
Set 3	193	38	20	23	10	8	2	1	2	1	1	0	0	0	0	0	0	0	0	0	0
Set 4	197	32	23	23	11	5	3	2	2	0	1	0	0	0	0	0	0	0	0	0	0
Set 5	192	38	21	22	11	7	3	1	3	1	0	0	0	0	0	0	0	0	0	0	0
EG 1a vs. 5																					
Set 1	167	15	8	9	11	11	6	5	8	2	14	11	6	7	2	6	1	1	1	0	8
Set 2	169	14	6	10	9	10	9	10	4	3	13	11	11	3	4	3	0	1	1	0	8
Set 3	169	16	11	2	8	11	9	5	7	5	14	10	9	5	4	2	3	0	0	1	8
Set 4	168	19	9	4	9	12	10	4	7	4	11	9	12	4	4	3	0	0	2	0	8
Set 5	167	15	11	7	9	10	9	7	4	3	13	11	11	3	5	3	1	1	1	0	8

Table 12 (cont'd)

Distribution of Item Alpha Values by
Reference/Focal (R/F) Comparison and Set

R/F Comp. and Set	Alpha Values																			
	<1.100	1.100-1.199	1.200-1.299	1.300-1.399	1.400-1.499	1.500-1.599	1.600-1.699	1.700-1.799	1.800-1.899	1.900-1.999	2.000-2.199	2.200-2.399	2.400-2.599	2.600-2.799	2.800-2.999	3.000-3.199	3.200-2.399	3.400-3.599	3.600-3.799	3.800-3.999	≥4.000
EG 1a vs. 6																					
Set 1	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
Set 2	185	17	24	16	13	11	7	6	3	4	3	4	2	2	1	0	0	0	0	0	1
Set 3	185	27	12	21	10	13	6	5	3	4	5	1	2	2	2	0	0	0	1	0	0
Set 4	187	23	16	20	9	16	5	3	4	2	4	4	1	4	0	0	0	0	0	1	0
Set 5	189	17	21	17	16	7	7	7	1	4	2	5	2	3	0	0	0	0	0	0	1
EG 1a vs. 7																					
Set 1	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
Set 2	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
Set 3	194	46	23	15	9	6	2	1	1	1	1	0	0	0	0	0	0	0	0	0	0
Set 4	200	38	33	12	6	2	4	1	1	1	1	0	0	0	0	0	0	0	0	0	0
Set 5	201	44	29	12	2	5	1	2	2	1	0	0	0	0	0	0	0	0	0	0	0
EG 1a vs. 1b																					
Set 1	297	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Set 2	299	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Set 3	298	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Set 4	299	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Set 5	299	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

*** Indicates a comparison was not conducted because the score distribution for the minority group did not meet the requirements for Minimum Score Group and Interval Size.

Table 13

Summary of Item Alpha Values Given in Table 12
by Reference/Focal (R/F) Comparison and Set

R/F Comparison and Set	Alpha Values				
	<1.100	1.100-1.499	1.500-1.999	2.000 2.999	≥3.000
EG 1a vs. 2					
Set 1	170	55	32	31	11
Set 2	173	54	31	31	10
Set 3	175	55	27	32	10
Set 4	171	57	29	32	10
Set 5	170	59	29	31	10
EG 1a vs. 3					
Set 1	185	81	27	6	0
Set 2	182	85	27	5	0
Set 3	183	88	23	5	0
Set 4	183	90	21	5	0
Set 5	183	89	21	6	0
EG 1a vs. 4					
Set 1	195	89	14	1	0
Set 2	194	88	17	0	0
Set 3	193	91	14	1	0
Set 4	197	89	12	1	0
Set 5	192	92	15	0	0
EG 1a vs. 5					
Set 1	167	43	32	40	17
Set 2	169	39	36	42	13
Set 3	169	37	37	42	14
Set 4	168	41	37	40	13
Set 5	167	42	33	43	14
EG 1a vs. 6					
Set 1	***	***	***	***	***
Set 2	185	70	31	12	1
Set 3	185	70	31	12	1
Set 4	187	68	30	13	1
Set 5	189	71	26	12	1
EG 1a vs. 7					
Set 1	***	***	***	***	***
Set 2	***	***	***	***	***
Set 3	194	93	11	1	0
Set 4	200	89	9	1	0
Set 5	201	87	11	0	0
EG 1a vs. 1b					
Set 1	297	2	0	0	0
Set 2	299	0	0	0	0
Set 3	297	1	0	0	0
Set 4	299	0	0	0	0
Set 5	299	0	0	0	0

*** Indicates a comparison not conducted because the score distribution for the minority group did not meet the requirements for Minimum Score Group Size and Interval Size.

BEST COPY AVAILABLE

Table 14

Post Hoc Analyses

Characteristics of Comparison Set 6

Set No. and R/F Comparisons	Mean Score Group Size		Cases Analyzed		Cases Excluded	
	Ref. Group	Focal Group	Ref. Group	Focal Group	Ref. Group	Focal Group
Set 6 (≥1 case)						
EG 1a vs. 2	263.3	62.9	18,693	4,469	5,142	2,017
EG 1a vs. 3	263.3	64.6	18,693	4,944	5,142	522
EG 1a vs. 4	263.3	23.6	18,693	1,679	5,142	325
EG 1a vs. 5	263.3	10.1	18,693	714	5,142	300
EG 1a vs. 6	263.3	5.3	18,693	378	5,142	108
EG 1a vs. 7	263.3	3.7	18,693	262	5,142	45
EG 1a vs. 1b	263.3	262.8	18,693	18,657	5,142	5,081

Table 15

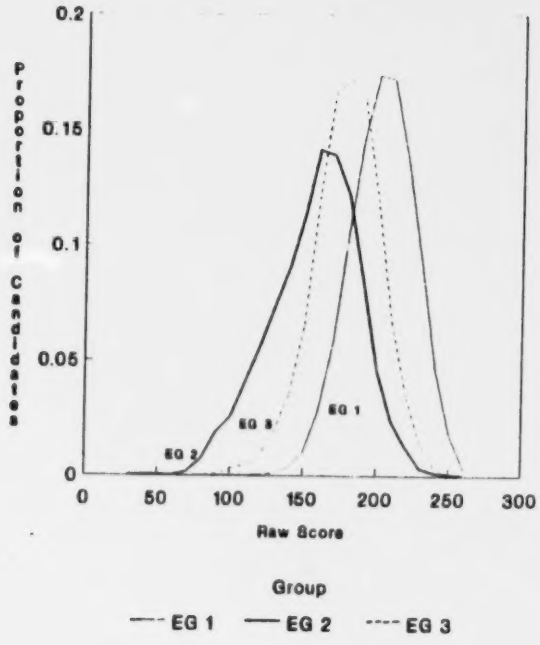
Post Hoc Analyses

Summary of Item Alpha Values
by Reference/Focal (R/F) Comparison and Set

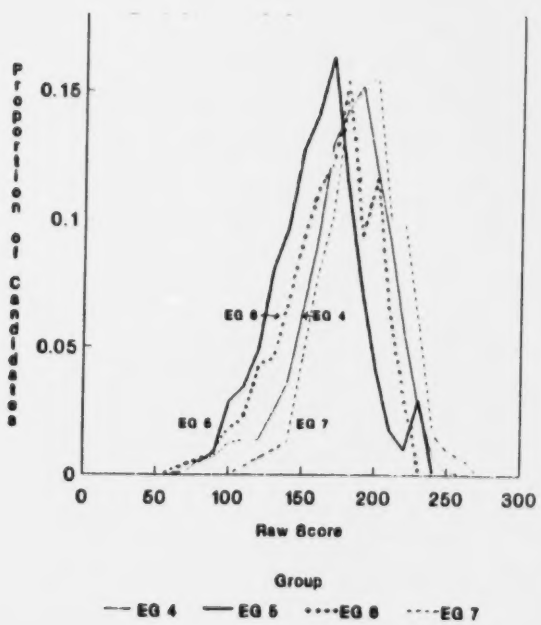
R/F Comparison and Set	Alpha Values				
	<1.100	1.100-1.499	1.500-1.999	2.000 2.999	≥3.000
EG 1a vs. 2 Set 6	170	64	26	29	10
EG 1a vs. 3 Set 6	185	88	21	5	0
EG 1a vs. 4 Set 6	198	86	14	1	0
EG 1a vs. 5 Set 6	169	43	33	42	12
EG 1a vs. 6 Set 6	187	72	27	12	1
EG 1a vs. 7 Set 6	201	90	7	1	0
EG 1a vs. 1b Set 6	299	0	0	0	0

Figure 1

Raw Score Distributions of Ethnic Groups 1, 2, and 3



Raw Score Distributions of Ethnic Groups 4, 5, 6, and 7



Appendix A

Results of *Post Hoc* MH Analyses for Minimum Score Group Size of One Case and Maximum Interval Size of One Raw Score Point (Set 6)

Table A.1

Post Hoc Analyses
MH Results by Reference/Focal (R/F), Set 6
Group Comparison and Set

R/F Comparison and Set	Alpha Mean	Alpha S.D.	Alpha Min.	Alpha Max.
EG 1a vs. 2 Set 6	1.195	.851	.157	8.240
EG 1a vs. 3 Set 6	1.053	.324	.411	2.300
EG 1a vs. 4 Set 6	1.036	.249	.439	2.012
EG 1a vs. 5 Set 6	1.293	1.098	.097	9.553
EG 1a vs. 6 Set 6	1.080	.460	.332	4.159
EG 1a vs. 7 Set 6	1.023	.215	.525	1.867
EG 1a vs. 1b Set 6	1.000	.025	.942	1.081

Table A.2

Post Hoc Analyses, Set 6
Distribution of Item Alpha Values by Reference/Focal (R/F) Comparison

R/F Comp. and Set	Alpha Values																				
	<1.100	1.100-1.199	1.200-1.299	1.300-1.399	1.400-1.499	1.500-1.599	1.600-1.699	1.700-1.799	1.800-1.899	1.900-1.999	2.000-2.199	2.200-2.399	2.400-2.599	2.600-2.799	2.800-2.999	3.000-3.199	3.200-2.399	3.400-3.599	3.600-3.799	3.800-3.999	≥4.000
EG 1a vs. 2																					
Set 6	170	20	11	17	16	9	5	4	4	4	11	7	5	4	2	1	2	1	1	2	3
EG 1a vs. 3																					
Set 6	185	29	21	20	18	13	4	4	0	0	5	0	0	0	0	0	0	0	0	0	0
EG 1a vs. 4																					
Set 6	198	33	22	20	11	5	5	3	1	0	1	0	0	0	0	0	0	0	0	0	0
EG 1a vs. 5																					
Set 6	169	14	7	9	13	10	7	9	4	3	13	11	9	5	4	2	1	1	0	0	8
EG 1a vs. 6																					
Set 6	187	18	29	10	15	10	5	6	4	2	4	2	3	3	0	0	0	0	0	0	1
EG 1a vs. 7																					
Set 6	201	48	25	13	4	3	0	3	1	0	1	0	0	0	0	0	0	0	0	0	0
EG 1a vs. 1b																					
Set 6	299	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0